DOCUMENT RESUME

ED 087 798 TM 003 421

AUTHOR Jacobs, Stanley S.
TITLE The Evaluation of the Culturally Different: Pre-School, Primary and Elementary Age Children.
PUB DATE 31 Oct 73
NOTE 22p.; Paper presented at the 4th Annual Convocation of the Northeastern Educational Research Association, Ellenville, N.Y.; October 31-November 2, 1973

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Children; Elementary School Students; Preschool Children; *Standardized Tests; *Student Evaluation; Syntax; *Test Bias; Test Construction; *Testing Problems; Test Reliability; Test Validity

ABSTRACT

This report asserts that the evaluation of young children can be viewed as a classic case of evaluation of the culturally different. Emphasized is the fact that not only are the majority of tests developed with an adult's perspective concerning adequacy of directions, items, and formats, but also the evaluation of the products is carried out within the context of adult experience and imperfect memory. Selected research literature concerning factors influencing the evaluation of young children is reviewed. Some new data is presented, including a syntactic analysis of verbal directions for children's tests and an analysis of the difficulty of directions read by children on other instruments. The normative equivalents of change-level scores on a number of standardized measures of cognitive variables are examined. The evidence supports the original conceptualization of cultural bias. (Author/NE)

ED 087798

TM 003 421

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The Evaluation of the Culturally Different:

Pre-school, Primary and Elementary Age Children

Stanley S. Jacobs
University of Pittsburgh

Paper Presented at the 4th Annual Convocation of the
Northeastern Educational Research Association
Fallsview Hotel Ellenville, New York

October 31 - November 2, 1973

The Evaluation of the Culturally Different: Pre-school, Primary and Elementary Age Children

Stanley S. Jacobs
University of Pittsburgh

Introduction

The "evaluation of the culturally different" in North America has typically been interpreted within the context of the evaluation of non-white, non-middle class minority groups, (e.g. Flaugher, 1970). Webster (1968), however, defines culture in a broader and psychometrically more relevant manner as "... a complex of typical behaviors or standardized social characteristics peculiar to a specific group, occupation or profession, sex, age, grade or social class ..." (p. 552). It is the basic premise of this paper that the probability that a test will be an inadequate (i.e. invalid) measure increases whenever a test is designed or developed by one group for use with another group, where there is (a) evidence that the groups differ with respect to variables related to test performance, and (b) an inability or unwillingness for the groups to communicate. Groups could differ on a number of cultural variables; this paper will concern itself only with age.

A great deal of attention has been paid, in recent years, to the problems which arise when middle-class whites design tests which are used with non-middle class, non-white groups. Most of the groups which feel that they have been or are being evaluated in an inadequate manner have spokesmen, a vocal constituency and legal recourse. There is, however, a large group of frequently-tested individuals who have had little to say about the manner or conditions of testing. This group is, of course, pre-school, primary and elementary age children.

The basic thesis of this paper is that there is a substantial body of evidence which, when integrated, indicates that children of this age bracket are being evaluated in a chronically inadequate fashion. Not only are the vast majority of tests developed with an adult's perspective concerning adequacy of directions, items, format, etc., but the evaluation of the products is similarly carried out, within the context of adult experience, imperfect memories, and attempts to introspect oneself into a child's shoes (see, for example, the evaluations of Hoeptner, Stern and Nummedal (1971) and Hoepfner (1970)). A number of current measurement texts mention or attempt to delineate important aspects of the problem. Stanley and Hopkins (1972), for example, devote a chapter to the psychological and cultural factors that influence performance on measures of cognitive variables; Cronbach (1970) also devotes a chapter to the problems associated with the measurement of ability in young children.

A number of studies will be reviewed, including those dealing with adequacy of directions, test and item format, response mode, coaching and test-taking skills, serial retesting and anxiety, in an attempt to delineate several dimensions of the problem.

The Adequacy of Test Directions Given Children

The evaluation of the adequacy of test directions given children is a complicated task. This stems from the fact that directions for most procedures common to pre-school and primary grade settings are read to students. This, of course, rules out any attempt at determining readability using general procedures such as the Dale-Chall formula (Dale and Chall, 1948a; 1948b) or procedures specifically developed for standardized tests (Forbes and Cottle, 1953).

The work of Chomsky (1969), investigating the acquisition of syntax in children, seems to bear on the problem, however, as does the work of Bormuth,

Carr, Manning and Pearson (1970), the work of Billington (1972) and Tatum's (1970) study. All showed a developmental sequence in children's ability to deal with certain definable linguistic structures. There does not seem to be any published work in the United States directly bearing on the relationship between the syntactic structure of test and item directions and test performance. However, in two studies conducted in Great Britain, Riley (1966) found 8 and 9 year-olds with below-average reading scores tended to score poorly on the verbal Essential Intelligence Test, and Cookson (1970) observed that children with reading age equivalents of less than 8.3 on the Schonell Graded Reading Test were apparently unable to understand the Junior Eysenck Personality Inventory.

A cursory examination of a few immediately available tests revealed use of syntactic structures difficult for children, in both test items and directions.

The confusion between "ask" and "tell" observed by Chomsky might lead to a child's misunderstanding such an item as "Do the children forget to ask you to play with them?"[1] from the California Test of Personality, Primary Form B. Potential problems with pronoun reference, observed by Chomsky, were noted in such examples as this quotation from the directions to the usage subtest of the California Achievement Tests, Lower Primary, Form W:

> Each sentence below has two words placed one above the other. You are to make an X on the one which you think is correct in each sentence.[2]

That the child understand these directions is important as he may otherwise assume the X should be placed on the incorrect word. If he is unable to cope with the indefinite "one," the subtest is not valid for him.

Billington's findings on the relationships of subordinate to main clauses

---

[1] Louis P. Thorpe, Willis W. Clark and Ernest W. Tiegs, *California Test of Personality, Primary Form B.* (California Test Bureau, 1942), p. 4.

[2] Ernest W. Tiegs and Willis W. Clark, *California Achievement Tests, Lower Primary, Form W,* (McGraw-Hill, 1957), p. 25.

were applied to the Gilmore Oral Reading Test, Forms C and D.[3] It was observed that paragraph D-4 contained five simple sentences and three complex sentences consisting of one main clause with a left-branched temporal clause. The corresponding paragraph of the parallel form C contained five simple sentences and two complex sentences, one a main clause with a right-branched adjective clause, and the other, a main clause with a right-branched noun clause from which was right-branched a temporal clause. Billington's findings suggest, although they do not show conclusively, that right-branching may be easier for children to understand than left-branching. Whether the three-clause complex sentence is equal to two two-clause sentences is an open question, as is the parallelism of Forms C and D.

The comparative-equal construction found to be particularly difficult for children by Bormuth, Carr, Manning and Pearson was frequently observed. Its use in directions is common and seems especially ill-advised. An example of its use in directions is in the Group Diagnostic Reading Aptitude and Achievement Tests:

> The teacher will show you some designs on a card. Study these designs until the teacher removes the card. Then draw as many of them as you can remember.[4]

This construction was frequently used in math subtests, but since it is part of the vocabulary of mathematics, this seems reasonable. One word problem of the Stanford Achievement Test, Primary I Battery, Form W, however, states:

> Make crosses on as many pints as you can fill if you empty the quart of milk into them.[5]

---

[3]John V. Gilmore and E.C. Gilmore, *Gilmore Oral Reading Test, Forms C and D* (Harcourt, Brace and World, 1968), p. 4.

[4]Marian Monroe and Eva Edith Sherman, *Group Diagnostic Reading Aptitude and Achievement Test* (C.H. Nelson Co., 1966), p. 11.

[5]Kelley L. Freeman, Richard Maddan, Eric F. Gardner and Herbert C. Rudman. *Stanford Achievement Test, Primary I Battery* (Harcourt, Brace & World, 1964), *Directions for Administering*, p. 22.

Although the comparative-equal concept may be necessary to the item, to include in its wording, also, a conditional clause at the end of a three clause construction seems unnecessarily complicated. Form X uses identical wording on a parallel item. Form Y, however, contains no item of comparable linguisitc complexity in the corresponding subtest.

In general, the questions asked on the tests examined appeared to be of the type termed "rote" by Bormuth, Carr, Manning and Pearson. However, in the third grade reading comprehension portion of the Iowa Tests of Basic Skills, Multilevel Edition for grades 3-9, the following actor-deleted, passively-transformed question was observed, "Why was the sign put on the door?"[6] Examination of the remaining "why" questions of the subtest revealed that, with one exception, all other "why" questions retained the active voice. The exception was in the seventh grade portion of the test. No reason for including this difficultly worded question in the third grade portion of the test is apparent.

Syntactic constructions not produced by all elementary school children were often observed in tests intended for them. Part F of the Social Adjustment Scale of the California Test of Personality, Primary Form B, was particularly loaded in this respect. Its items included the following:[7]

> Is there a nice group of children of your own age in your neighborhood with whom you play?

Notice that this is a there-insertion sentence containing a transformation-produced nominal used as the object of a preposition.

> Are some of the people near your home so mean that you like to do things to make them angry?

Here are, not one, but two adverbial infinitives.

---

[6] E.F. Lindquist and A.M. Hieronymus, *Iowa Tests of Basic Skills, Form I*, Multilevel Edition (State University of Iowa, 1955), p. 8.

[7] *California Test of Personality, Primary Form B*, p. 14.

Are conditions in your neighborhood as good as you would like to have them?

This sentence uses the comparative-equal construction as well as an adverbial infinitive. Children's understanding of the subjunctive mood, which it also employs, has not yet been investigated. Although others of the twelve parts of the California Test of Personality contain items with difficult wordings, none seemed as linguistically difficult as the one in which these items appear.

Measurement procedures designed for the upper elementary grades do require that the child read and understand test and item directions. An analysis of seven commonly used batteries which were available for analysis (see Table I) was carried out using the Dale-Chall formula with three 100 word samples, drawn from the beginning, middle and end of the total material read by the test-taker.

Table I

Dale-Chall Scores for Selected Intermediate Level Batteries and Subtests

Test Level / Form / Sub-Test	Stanford Intelligence II (W)	Stanford Intelligence I (X)	Metropolitan Intelligence (B)	Metropolitan Intelligence (G)	California Achievement Test 4,5,6, 7,8 (E)	Iowa Test of Basic Skills 3-9 (I)	Comprehensive Test of Basic Skills 6,7,8, (A)	$\bar{X}$ Subtest Score
Vocab/Wd Mng	4.28	4.28	5.79	5.26		5.10		4.94
Par Rdng	5.19	5.19	4.90	4.52		5.09	5.34	5.04
Spelling	4.76	4.76	5.04	6.03				5.13
Lang	5.39	5.97	5.42	4.67		5.57	5.59	5.43
Math	4.69	4.71	5.67	5.09	5.19	5.44	5.44	5.18
Soc St	4.91	4.57	4.52	5.09		5.12		4.84
Science	4.19	4.18	4.67	5.27				4.58
Wrd Stdy Sk		4.84					5.68	5.26
Memory					4.27			4.27
Spatial					4.41			4.41
Logic					4.98			4.98
Verbal					4.73			4.73
$\bar{X}$ test score	4.77	4.81	5.14	5.13	4.72	5.26	5.51	

The results shown in Table I indicate that the range of readability on intermediate level measures is from approximately the 4th to the 6th grade level, with a mean for tests of close to 5th grade level. When tests containing such directions are given to 4th (or even many 5th) grade children, the reading difficulty of the directions may prevent their comprehension of what is desired or required. This is very probably a problem common to most tests or batteries designed for a span of several grades at the elementary level.

The Effect of Response Mode on Performance

The introduction of machine scorable answer sheets in educational measurement should have been welcome, preventing, among other things, error rates of the magnitude documented by Phillips and Weathers (1958), who found that 28% of a sample of teacher-scored standardized tests contained errors. They should have been welcomed as an economy, allowing the use of relatively more expensive batteries printed in booklets which would be reusable over several years. Instead, there was suspicion from the outset that primary children could not contend with the use of a separate answer sheet.

Cashen and Ramseyer (1969) found differences between the booklet and answer sheet response modes (favoring the former) to be significant at grades 1 and 2, but not at grade three. Gaffney and Maguire (1971), presented evidence that, even with a brief orientation to the use of a separate answer sheet and a short practice session children below grade four seem to have difficulty. They also found that with only orientation and no practice, children in grade five or below had difficulty with the response mode. Ramseyer and Cashen (1971) attempted to develop a training procedure to enable first and second graders to respond using a separate answer sheet. A 20 minute practice session involving an introduction to the use of separate answer sheets was provided experimental Ss. No significant gains due to training were reported, indicating that even after an

orientation procedure, first and second grade students were unable to contend with the demands of separate answer sheets. Solomon's (1971) study with inner-city, culturally deprived fourth graders indicated that there were no significant differences between test booklet and separate answer sheet formats. However, there are two points of criticism in this study: six of a total of 45 Ss were dropped from the separate answer sheet condition because they "... failed to follow instructions." (p. 290), and the children are described as having "... had previous experience with machine scorable answer sheets." His conclusions ought not to be taken seriously. Muller, Calhoun and Orling (1972), in an experiment involving third, fourth and sixth grade students, found the number of marking errors for a separate answer sheet group to be about three times that for a group responding in a test booklet at each of the three grade levels, in response to specially developed items which were assumed to be common knowledge to children at those levels.

To summarize, it appears that the use of separate answer sheets is inadvisable with children below grade six. The ability to contend with the demands of this response mode may be teachable, and is probably a function of a number of "subject variables." This is still a largely unanswered question.

The Effects of Coaching, Practice and Test-Wiseness

Surprisingly little data of a trustworthy nature has been generated on the question of test familiarization with students in the United States. One of the better, though somewhat dated, summaries of the British experience was reported by Vernon (1954), after a symposium dealing with the question. To briefly summarize their conclusions:

a) It appears that non-verbal test material is more affected than verbal.

b) It appears the typical gain is from .4 to 1 s.d., depending upon the difficulty level of the test

c) The more naive students are, the greater will be the observed gains.

d) The most efficient and effective procedure seems to be a combination of coaching and actual timed practice with the test or tests of concern.

e) An important possible effect which had (at that time) not been investigated was the reduction of debilitating anxiety.

Slakter, Koehler and Hampton (1970) reported a study which purported to trace the development of test-wiseness over the grades 5 to 11. Although the results are interpreted as indicating a linear trend in the development of test-wiseness over the range of grades studied, several points should be stressed.

a) The data appeared extremely unreliable for Ss below the ninth grade; an estimate of the reliability of the TW measure for grades 5-8 would be .50.

b) Although it is claimed that grade effects were significant, an examination of the data indicates either trivial differences between adjacent grades, or a deterioration in performance at higher grades (see especially their Figure 1 for grades 5-7 and 9-11).

c) The same test-wiseness measures were used for grades 5-11. To claim that a trend in the data indicates a developmental trend in test-wiseness behavior reveals startling faith concerning the range of applicability of the measures used in the study.

Mann, Taylor, Proger, Dungan and Tidey (1970) investigated the effects of simple serial retesting with a sample of 7th graders, and found that significant gains did occur over the four trials given Ss with the greatest gains from first to second trial. No instructive feedback was provided students, nor was there any attempt at test-relevant or content-independent coaching. Although it was demonstrated that test anxiety level was independent of gains over trials, there were no data concerning the mechanism producing the observed gains.

Diamond and Evans (1972), in an investigation of the cognitive correlates of test-wiseness with sixth grade Ss, reported the median part-whole intercorrelation for five scales designed to tap different facets of testwiseness was .58; the correlation with Lorge-Thorndike IQ (total) was .49, with Iowa Test of Basic Skills Vocabulary subtest (a good "rough and ready" IQ measure) .55. The authors conclude that testwiseness is "not a pervasive skill" and "these responses have little relationship to a student's general cognitive ability." (p. 150). I would say that their own data contradicts their conclusions, and, at the very least, any conclusions based on 6-item scales are probably unreliable.

Callenbach (1973) has reported what appears to be the only recent study which seems sensitive to the conclusions and recommendations of the Vernon symposium, made almost 20 years ago. Twenty-four relatively naive second grade Ss received eight 30-minute periods of deliberate instruction and practice in content-independent test-taking skills over a four week period. In a comparison with control Ss, it was found that the treatment resulted in a significant gain in the experimental group of about .75 s.d., with a significant difference between experimental and control groups' posttest scores. Unfortunately (for a clear-cut interpretation) an analysis of the gains made by control Ss indicated a significant gain in that group as well. One is left wondering what the combination of practice and instruction (as recommended by Vernon) would produce, in a study designed to estimate the effects of the factors in a planned, rather than post-hoc, fashion.

One is drawn to the conclusion arrived at by Vernon almost 20 years ago -- that a combination of instruction and practice can have a statistically and practically significant impact on naive Ss. When we are discussing elementary school children, probably the vast majority of students in the lower grades ought to be regarded as test-naive. It would seem a deliberate testing-instruc-

tional effort, paralleling Boehm's Test of Basic Concepts, (Boehm, 1969) but in the area of test-taking skills and concepts, would be one approach to the removal of what may be a substantial source of error variance in young children's data.

Some Additional Evidence That All is Not Well With Standardized Tests

Hoepfner and Doherty (1973) summed up and analyzed ratings which had been assigned the tests produced by seven major American test publishers. The MEAN evaluation system (Hoepfner, 1970; Hoepfner, et al., 1971), assessing measurement validity, examinee appropriateness, administrative usability and normed technical excellence, indicated the majority of tests were good with respect to administrative usability, but fair to poor in all other respects. In other words, it seems that a great number of tests are available which make it very easy to generate meaningless data.

Is this a fair indictment, you may well ask. To gain some insight into the problem, a small study was carried out to determine the normative equivalents of chance-level performance on an available set of standardized tests of cognitive variables, designed for elementary school children. No claim is made for a random sample of measures, but the sample was drawn only on the basis of amenability to analysis, so deficiencies noted may well be symptomatic of the entire population of interest. Also, in discussing the notion of chance-level performance, we are actually speaking as though children responded in a random fashion to test items. I am reasonably sure this is usually not the case. However, for reasons already discussed (e.g. inability to comprehend verbal or written instructions, inability to utilize the response format of a test, naivete with respect to certain test-taking skills, debilitating anxiety), the testing experience may be so threatening or confusing to a child that the end result is the same; performance at or about the level that random responses would produce.

The question remains -- what is the incidence of the problem? A sample of 13 tests and batteries from files maintained by the Office of Measurement and Evaluation at the University of Pittsburgh was identified, and for each scale, subtest or test for which analysis was possible, an expected chance score and standard deviation for chance scores was calculated, using Gulliksen's formulae (1950, p. 263), which are, respectively:

$$M_c = n/c \text{ and } s_c = \frac{\sqrt{n(c-1)}}{c}$$

where M_c is the chance score

n is the number of test items

c is the number of options per item, and

s_c is the standard deviation of chance scores.

Gulliksen further recommends that scores lower than M_c+2s_c be regarded as not indicating knowledge of the variables under consideration, i.e. that scores lower than M_c+2s_c not be taken seriously.

Accordingly, Table 2 was constructed, in which normative equivalents (percentile rank, grade equivalents or IQ) are given for both M_c and M_c+2s_c.

Table 2

Chance Scores, Their Respective Standard Deviations, and Their Normative Equivalents on a Sample of Standardized Tests Designed for Pre-school, Primary and Elementary Pupils

Test and Intended Grade Level(s)	Chance Score (Mc)*	Chance S.D. (s_c)	% ile Rank	Grade Equiv	IQ	$M_c + 2s_c$	% ile Rank	Grade Equiv	IQ
American School Achievement Tests: Primary Battery II Grades 2 and 3									
Sentence & Word Meaning	$7.5\frac{7}{8}$	2.4	--	$\frac{1.7}{1.8}$	--	$\frac{12}{13}$	--	$\frac{2.4}{2.6}$	--
Paragraph Meaning	$7.5\frac{7}{8}$	2.4	--	$\frac{1.7}{1.8}$	--	$\frac{12}{13}$	--	$\frac{2.3}{2.5}$	--
Arith. Computation	10	2.7	--	2.0	--	15	--	2.6	--
Arith. Problems	3	1.5	--	1.9	--	6	--	2.9	--
Language Usage	6	2.8	--	1.4	--	12	--	1.8	--
Spelling	$7.5\frac{7}{8}$	2.4		$\frac{1.6}{1.7}$	--	$\frac{12}{13}$	--	$\frac{2.1}{2.2}$	--
American School Achievement Tests: Intermediate Battery Grades 4-6									
Sentence & Word Meaning	10	2.7	--	2.3	--	15	--	2.9	--
Paragraph Meaning	10	2.7	--	2.6	--	15	--	3.4	--
Arith Computation	10	2.7	--	3.9	--	15	--	4.8	--
Arith Problems	5	1.9	--	3.9	--	9	--	4.7	--
Spelling	$12.5\frac{12}{13}$	3.0	--	$\frac{3.4}{3.5}$	--	$\frac{18}{19}$	--	$\frac{4.0}{4.1}$	--
Social Studies	10	2.7	--	4.4	--	15	--	5.3	--
Science	10	2.7	--	4.5	--	15	--	5.4	--

*When scores in this and the $M_c 2s_c$ column are not whole numbers they are entered in their respective columns in the form Xa/Xb, representing the score above and below the computed value.

Test and Intended Grade Level(s)	Chance Score (Mc)	Chance SD (s_c)	% ile Rank	Grade Equiv	IQ	$M_c + 2s_c$	% ile Rank	Grade Equiv	IQ
California Test of Mental Maturity, 1957 S-Form Grades K-1									
Spatial Relationships	5.8 5/6	2.5	40/60	--	--	10/11	99/99	--	--
Verbal Concepts	6	3.0	2	--	--	12	70	--	--
California Test of Mental Maturity, 1963 S-Form Level-0, Grades K-Low 1									
Factor I	7.6 7/8	2.3	69/79	--	--	12/13	97/99	--	--
Factor II	4	1.8	38	--	--	8	98	--	--
Factor III	6	2.0	24	--	--	8	46	--	--
Factor IV	2	1.2	31	--	--	4	69	--	--
Total Test	19/20	7.3	42/46	--	--	32/33	92/93	--	--
California Test of Mental Maturity, 1957 S-Form Elemen. Gr. 4-8									
Factor I	13.8 13/14	3.9	20/20	--	--	21/22	50/60	--	--
Factor II	8.8 8/9	3.5	10/10	--	--	15/16	40/40	--	--
Factor IV	12.5 12/13	3.0	80/80	--	--	18/19	95/95	--	--
Cognitive Abilities Test, Primary 1/Form 1, Ages 5-0 to 8-0	16	5.3	2	--	68	26	16	--	84
Kuhlmann-Anderson Test, 1964 Booklet A, Grade 1-?	9.5 9/10	3.5	4/4	--	71/72	16/17	9/9	--	79/80

Test and Intended Grade Levels	Chance Score (M_c)	Chance SD (s_c)	% ile Rank	Grade Equiv	IQ	$M_c + 2s_c$	% ile Rank	Grade Equiv	IQ
Otis-Lennon Mental Ability Test, Elementary I Level Gr. Mid. Gr. 1-3	20	3.8	38	--	95	28	69	--	108
Otis-Lennon Mental Ability Test, Primary I Level, Grade: Last half K.	13.8 13/14	3.2	16/19	--	84/86	19/20	38/43	--	95/97
SRA Primary Mental Abilities Test, Gr. K-1									
Verbal Meaning	12.5 12/13	3.0	--	--	80/83	18/19	--	--	93/97
Perceptual Speed	7	2.3	--	--	100	11	--	--	100
Number Facility	0	0.0	--	--	--	0	--	--	--
Spatial Relations	3	1.5	--	--	67	6	--	--	83
Total Test	22/23	6.8	--	--	83/87	35/36	--	--	93/93
SRA Primary Mental Abilities Test, Gr. 2-4									
Verbal Meaning	15	3.4	--	--	68	22	--	--	76
Spatial Relations	6.8 6/7	2.2	--	--	81/85	10/11	--	--	92/95
Number Facility	1.2 1/2	0.9	--	--	74/78	3/4	--	--	81/83
Perceptual Speed	8.3 8/9	2.6	--	--	83/84	13/14	--	--	92/95
Total Test	30/33	9.7	--	--	50/51	48/51	--	--	66/69
Stanford Achievement Test Primary I Gr. Md. 1-Beg 2									
Word Meaning	8.8 8/9	2.6	8/12	1.1/1.2	--	13/14	32/40	1.4/1.5	--
Paragraph Meaning	9.5 9/10	2.7	22/36	1.4/1.5	--	14/15	50/50	1.6/1.6	--

Test and Intended Grade Level(s)	Chance Score (M_c)	Chance SD (s_c)	% ile Rank	Grade Equiv	IQ	$M_c + 2s_c$	% ile Rank	Grade Equiv	IQ
Stanford Achievement Test Primary I, Gr. Mid. I-Beg 2									
Vocabulary	13	2.9	12	1.3	--	19	54	1.7	--
Word Study Skills	18.6 $\frac{18}{19}$	3.5	$\frac{6}{11}$	$\frac{1.1}{1.2}$	--	$\frac{25}{26}$	$\frac{20}{30}$	$\frac{1.3}{1.4}$	--
Arithmetic	9.5 $\frac{9}{10}$	3.1	$\frac{1}{4}$	$\frac{1.0}{1.1}$	--	$\frac{15}{16}$	$\frac{12}{12}$	$\frac{1.2}{1.2}$	--
Stanford Achievement Test Primary II Battery, Gr. Mid 2-3									
Word Meaning	9	2.6	14	1.8	--	14	42	2.5	--
Paragraph Meaning	15	3.4	16	1.9	--	22	38	2.4	--
Science & Soc. Study Concepts	12.6 $\frac{12}{13}$	2.9	$\frac{18}{18}$	$\frac{1.6}{1.8}$	--	$\frac{18}{19}$	$\frac{56}{62}$	$\frac{2.7}{2.9}$	--
Word Study Skills	18.8 $\frac{18}{19}$	5.1	$\frac{8}{11}$	$\frac{1.5}{1.6}$	--	$\frac{28}{29}$	$\frac{36}{42}$	$\frac{2.3}{2.4}$	--
Language	28.2 $\frac{28}{29}$	3.7	$\frac{22}{28}$	$\frac{2.2}{2.3}$	--	$\frac{35}{36}$	$\frac{54}{60}$	$\frac{2.7}{2.8}$	--
Arithmetic Concepts	9.7 $\frac{9}{10}$	3.8	$\frac{11}{16}$	$\frac{1.7}{1.9}$	--	$\frac{17}{18}$	$\frac{42}{48}$	$\frac{2.6}{2.7}$	--

As is distressingly obvious, for several tests, a substantial proportion of the norming sample (and, by inference, children like them) performed in a manner which produced scores similar to what random responding would have produced.

I feel that this is a rather strong indictment of the type of data typically collected with groups of elementary-age youngsters. I think it's past time that researchers in education and psychology stopped deluding themselves and others with rationalizations to account for the common low reliabilities and predictive validities for data collected from young children. Irregularities in human development, differences in test content and constructs measured, and the discrepancy between individual growth curves and a curve based on group averages are common "explanations" for the unstable data. These and other, factors may well exert an influence on the data.

I suspect, however, that a fundamental reason is that, for large numbers of primary and elementary age children, the manner and type of testing done is inappropriate. I further suspect that the problem is one with at least a partial solution:

a) Develop a measure or measures of "readiness" for standardized testing,

b) Develop training experiences to prepare children for standardized testing and

c) Eliminate the more inadequate tests from consideration in a testing program.

This last consideration might involve an examination of tests, evaluations in sources such as Buros' *Mental Measurements Yearbooks*, journals and other published sources, as well as personal evaluations based on, for example, the latest draft of the APA-AERA-NCME *Standards for Development and Use of Educational and Psychological Tests*. Systematic observation and evaluation of children's behavior on standardized tests might detect the operation of some

of the problems which lead to data like that presented in Table 2. This would seem a necessary first step in the development of either a) or b) above.

REFERENCES

Billington, D.R. The effects of subordinate clause type, position and number on the development of children's perceptions of the relationships among clauses in complex sentences. *Dissertation Abstracts*, 1972, 32, 5032.

Boehm, A.E. *Manual, The Boehm Test of Basic Concepts*. New York: Psychological Corporation, 1969.

Bormuth, J.R., Carr, J., Manning, J. and Pearson, D. Children's comprehension of between-and-within sentence syntactic structures. *Journal of Educational Psychology*, 1970, 61, 349-357.

Callenbach, C. The effects of instruction and practice in content-independent test-taking techniques upon the standardized reading test scores of selected second-grade students. *Journal of Educational Measurement*, 1973, 10, 25-30.

Cashen, V.M. and Ramseyer, G.C. The use of separate answer sheets by primary age children. *Journal of Educational Measurement*, 1969, 6, 155-158.

Chomsky, C. *The Acquisition of Syntax in Children from 5 to 10*. Cambridge, Massachusetts: MIT Press, 1969.

Cookson, D. Study of difficulties in reading and understanding the Junior Eysenck Personality Inventory. *British Journal of Educational Psychology*, 1970, 40, 8-14.

Cronbach, L.J. *Essentials of Psychological Testing* (3rd edition). New York: Harper and Row, 1970.

Dale, E. and Chall, J. A formula for predicting readability. *Educational Research Bulletin*, 1948a, 27, 1-28.

Dale, E. and Chall, J. A formula for predicting readability: instructions. *Educational Research Bulletin*, 1948b, 27, 37-54.

Diamond, J.J. and Evans, W.J. An investigation of the cognitive correlates of test-wiseness. *Journal of Educational Measurement*, 1972, 9, 145-150.

Flaugher, R.L. *Testing Practices, Minority Groups and Higher Education: A Review and Discussion of the Research*. Princeton: Educational Testing Service, 1970 (RB-70-41).

Forbes, F.W. and Cottle, W.C. A new method for determining readability of standardized tests. *Journal of Applied Psychology*, 1953, 37, 185-190.

Gaffney, R.F. and Maguire, T.O. Use of optically scored test answer sheets with young children. *Journal of Educational Measurement*, 1971, 8, 103-106.

Gove, P.B. (ed.) *Webster's Third New International Dictionary*. Springfield, Massachusetts: G & C Merriam Co., 1968.

Gulliksen, H. *Theory of Mental Tests*. New York: Wiley, 1950.

Hoepfner, R. (ed.) *Elementary School Test Evaluations*. Los Angeles: Center for the Study of Evaluation, University of the City of Los Angeles, 1970.

Hoepfner, R. and Doherty, W.J. Priorities of test publishers. *Journal of Educational Measurement*, 1973, 10, 85-93.

Hoepfner, R., Stern, C. and Nummedal, S.G. *Preschool and Kindergarten Test Evaluations*. Los Angeles: Center for the Study of Evaluation, University of the City of Los Angeles, 1971.

Mann, L., Taylor, R.G. Proger, B.B., Dungan, R.H. and Tiday, W.J. The effect of serial retesting on the relative performance of high- and low-test anxious seventh grade students. *Journal of Educational Measurement*, 1970, 7, 97-104.

Muller, D., Calhoun, E. and Orling, R. Test reliability as a function of answer sheet mode. *Journal of Educational Measurement*, 1972, 9, 321-324.

Phillips, B.N. and Weathers, G. Analysis of errors in scoring standardized tests. *Educational and Psychological Measurement*, 1958, 18, 563-567.

Ramseyer, G.C. and Cashen, V.M. The effect of practice sessions on the use of separate answer sheets by first and second graders. *Journal of Educational Measurement*, 1971, 8, 177-181.

Riley, C.S. Relationship between reading ability and verbal intelligence test performance. *British Journal of Educational Psychology*, 1966, 36, 117.

Slakter, M.J., Koehler, R.A. and Hampton, S.H. Grade level, sex and selected aspects of test-wiseness. *Journal of Educational Measurement*, 1970, 7, 119-122.

Solomon, A. The effect of answer sheet format on test performance by culturally disadvantaged fourth grade elementary school pupils. *Journal of Educational Measurement*, 1971, 8, 289-290.

Standards for Development and Use of Educational and Psychological Tests. (3rd draft). Washington: American Psychological Association, 1973.

Stanley, J.C. and Hopkins, K.D. *Educational and Psychological Measurement and Evaluation*. Englewood Cliffs: Prentice-Hall, 1972.

Tatum, S.W. Reading comprehension of materials written with select oral language patterns: a study at grades two and four. *Reading Research Quarterly*, 1970, 5, 402-426.

Vernon, P.E. Symposium on the effects of coaching and practice in intelligence tests. *British Journal of Educational Psychology*, 1954, 24, 57-63.